FragPELE: Dynamic ligand growing within a binding site. A novel tool for hit-to-lead Drug Design.

Carles P. Lopez[†]*, Daniel Soler*[§]*, Robert Soliva*[§]*, and Victor Guallar*[†¥]***

[†]Life Sciences Department, Barcelona Supercomputing Center (BSC), Barcelona, 08034, Spain

[§]Nostrum Biodiscovery, Carrer Jordi Girona 29, Nexus II D128, 08034 Barcelona, Spain

[¥]ICREA: Institució Catalana de Recerca i Estudis Avançats, Passeig Lluís Companys 23, 08010 Barcelona, Spain

Keywords: Drug Design, Ligand Growing, Hit to Lead, PELE, Computational Chemistry

The early stages of drug discovery rely on hit-to-lead programs, where initial hits undergo partial optimization to improve binding affinities for their biological target. This is an expensive and time-consuming process, requiring multiple iterations of trial and error designs, an ideal scenario for applying computer simulation. However, most state-of-the-art modeling techniques fail to provide a fast and reliable answer to the induced-fit protein-ligand problem. To aid in this matter, we present FragPELE, a new tool for *in silico* hit-to-lead drug design, capable of growing a fragment from a bound core while exploring the protein-ligand conformational space. We tested the ability of FragPELE to predict crystallographic data, even in cases where cryptic sub-pockets

open due to the presence of particular R-groups. Additionally, we evaluated the potential of the software on growing and scoring five congeneric series from the 2015 FEP+ dataset, comparing them to FEP+, SP and Induced-Fit Glide, and MMGBSA simulations. Results show that FragPELE could not only be useful for finding new cavities and novel binding modes in cases where standard docking tools can not, but also to rank ligand activities in a reasonable amount of time and with acceptable precision.

INTRODUCTION

Hit to lead (H2L) efforts in early drug discovery pursue the optimization of potency as well as other properties such as solubility, absorption, chemical stability, and toxicity, to name a few.[1] Regarding potency, the most natural way for medicinal chemists to optimize a new chemical entity is to gain further interactions with the targeted receptor by growing the initial seed compound. Generally, hits are grown into leads by gaining electrostatic and hydrophobic interactions with their receptors. It has been described that the evolution of a hit to a clinical candidate yields, on average, gains of molecular weight (MW) ca. 85Da and increasing structural complexity.[2]

The “growing” strategy is especially relevant within the fragment-based drug design (FBDD) paradigm, where the starting points are typically low MW compounds with marginal potency on the receptor (typically double-digit μM). Fragments can either be linked or grown so that their potencies can be brought down to the nM level. However, drug-hunting teams applying FBDD primarily resort to fragment growing with the help of a series of detailed experimental 3D information, above all X-ray.[3,4]

Decorating a known chemical core on one or more points (R-groups) is not an obvious task, even when access to rich structural 3D information is available. Receptor motion both at the

side-chain and backbone level and the role of solvent are the main factors hampering a straightforward estimation on what type of chemical groups should be added to an R-group. Moreover, binding sites might undergo larger rearrangements as a response to ligand probing and open up cryptic sub-pockets. A prototypical example of the latter behavior can be seen by comparing the structure of epidermal growth factor receptor (EGFR) when bound to the first-generation inhibitor Gefitinib[5] vs. second-generation inhibitor Lapatinib.[6] The structure of the latter would have never been guessed by inspecting the X-ray of the EGFR-Gefitinib complex as significant rearrangements are needed to accommodate a much bulkier R-group beyond the gatekeeper residue.

Computational modeling is now a mainstream technology with applications in all early discovery tasks, from receptor prioritization, binding site detection and druggability analysis, to virtual screening and lead optimization. Structure-based in silico R-group exploration is not an exception, which can be handled by a variety of techniques. A quick and useful approach is to build a focused combinatorial library around the known core whose X-ray crystal structure is available and dock the whole series on the latter.[7] This can be done in an unguided fashion or directed by pharmacophoric or positional restraints to keep the core at the known position. A similar approach is to use generative models such as recurrent neural networks (RNN) or variational autoencoders (VAE)[8] to grow drug-like fragments onto an initial compound to dock them subsequently. An additional approach is to use de novo algorithms with the seed core placed at the binding site. Plenty of programs and modeling platforms offer such strategies. Among the most thoroughly tested de novo algorithms we find LigBuilder2.0[9] and Autogrow3.0.[10] Some of them are clever enough to incorporate synthetic feasibility such as SYNOPSIS[11] and NAOMINext.[12] However, most of them are based on a rigid representation of

the binding site,[13] which severely limits the generation of high-quality geometries and binding energies for the grown derivatives. Some attempts at incorporating receptor flexibility when expanding R-groups have been developed. For instance, induced-fit Glide, LEA3D,[14] SkelGen,[15] and OpenGrowth,[16] where ligand and local side chains are sampled to take into account the dynamics of the binding site.

R-group exploration can be also carried out with a variety of molecular dynamics (MD) based techniques that are generically used for studying a series of compounds against the same receptor. These approaches, although computationally more expensive, allow to partially capture binding site flexibility. Some approaches worth mentioning are Linear interaction approximation,[17] MMPB(SA) and MMGB(SA).[8] However, they are not specifically designed for R-group growth of a seed molecule placed at a binding site. The most rigorous but computationally expensive MD-based techniques with R-group applicability are based on alchemical free energy methods.[19,20] By way of thermodynamic cycles, they mutate a starting structure to a final structure providing a rigorous relative change in binding free energy.[19,21] While this technology is clearly among the most reliable for estimating binding free energies, it comes, however, with a considerable computational cost.

A dynamic exploration of a growing molecule in a binding site can also be explored with Monte Carlo (MC) algorithms, in fact, the program BOMB has been successfully applied to a variety of pharmaceutically relevant targets. This technique grows several fragments by replacing substituents of ligands that can be previously placed in the binding site or isolated. Conformational searches are performed for each molecule and conformers are optimized, evaluating then the lowest-energy conformers with quick scoring functions.[22]

MC codes have been widely used to achieve faster explorations of the protein-ligand conformational space. For instance, BLUES apply a nonequilibrium candidate Monte Carlo (NCMC) method to improve sampling of ligand binding modes.[23] A similar approach is represented by our platform PELE (Protein Energy Landscape Exploration). It implements a heuristic MC procedure including complex protein structure prediction techniques in each MC move.[24] Successful applications in drug design include poses prediction,[25] binding path sampling[26,27] and binding free energy prediction when coupled to MSM[28,29]. Furthermore, we recently introduced an adaptive sampling procedure, AdaptivePELE, improving the sampling performance (speed) by an order of magnitude.[30] Overall, the platform is very efficient at sampling rugged potential energy surfaces and avoiding entrapment in local minima when reproducing, for instance, ligand-receptor induced-fit complex formation.

Given its sampling efficiency, it was decided to test whether PELE could be adapted to R-group growth in the context of H2L. Herein, we describe a new protocol called FragPELE, based on gradually growing an R-group for a bound ligand in a series of steps. It has been devised to fall midway between the expensive alchemical free energy methods and the quick standard docking-based techniques. We report a first benchmark where we reveal that FragPELE is not only able to generate good geometries even for complex cases involving significant induced-fit effects, but can also yield binding energy estimates that are not far off from those obtained by the more expensive alchemical free energy methods. Interestingly, FragPELE shows good performance at locating cryptic sub-pockets in a known binding site. Overall, the methodology seems to be a promising strategy to explore R-group growth while capturing induced fit effects and represents a further addition to in silico H2L techniques.

METHODS

Computational method. FragPELE is based on the PELE software, which combines an MC stochastic approach with protein structure prediction techniques. Briefly, in PELE each MC step is composed by three processes: 1) initial perturbation, where the ligand is randomly rotated and translated and the atoms of the backbone of the protein are displaced through the use of normal modes; 2) side-chain sampling, where an experimental rotamer library is used to reposition side chains in response to the initial perturbation and; 3) final minimization, where the overall structure is relaxed. More details on the PELE method can be found on our recent book chapter.[31] In addition, FragPELE also uses BioPython[32] and ProDy[33] external libraries.

The software is intended to automatically grow one or more fragments onto different hydrogens of the same scaffold employing consecutive steps (epochs), following the technology recently introduced in AdaptivePELE.[30] To efficiently sample re-arrangement of the system as the fragment is grown, several independent PELE simulations are run at each epoch, adding extensive side-chain sampling, normal modes backbone sampling and minimization procedures. The overall method, shown in Figure 1, involves: preparation (1), fragment linkage (2), fragment reduction (3), fragment growing (4) and sampling/scoring (5).

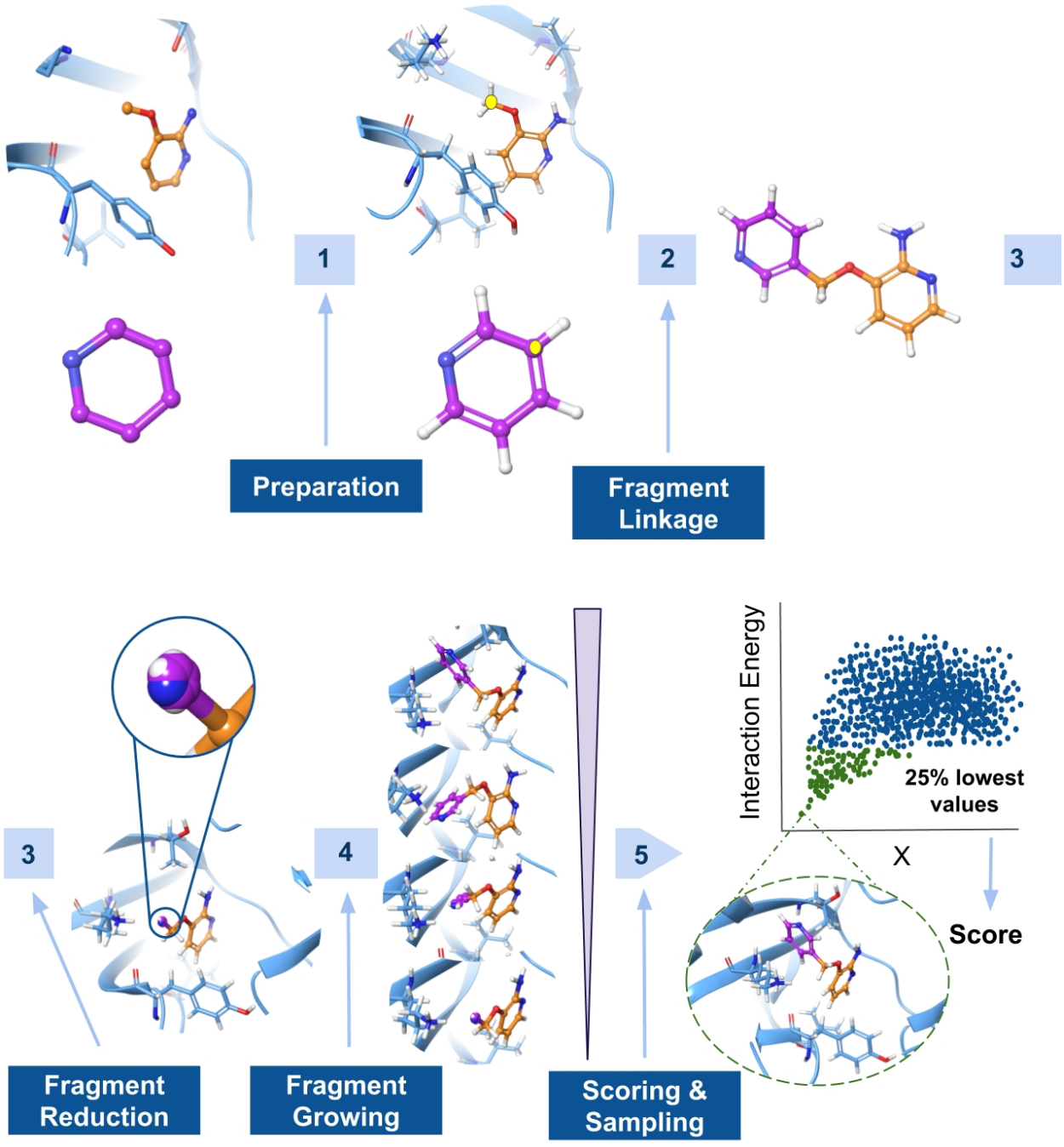

Figure 1. Schematic overview of the procedure to run Frag-PELE. In yellow we underline the heavy atoms that will be connected in the fragment linkage.

1 - Preparation. FragPELE is intended to automatically grow one or more fragments onto the same scaffold. For this, a previously prepared protein-ligand complex must be provided in PDB format (the complex can be an X-ray co-crystal structure, a docked pose for a chemical core generated by docking, etc.). Protonation and structural issues such as missing loops, missing atoms, and multiple occupancies must be solved previously. Metal ions, cofactors, and structural waters can be included as part of the receptor. All systems were prepared with Schrödinger's Protein Preparation Wizard[34] including H-bond optimization using PROPKA[35] at pH 7. The fragments to be grown must be provided in separate PDB files with the correct protonation states, together with a tabular input file where the user must specify the path to the fragment files and the atoms to be linked.

2 - Fragment linkage. The purpose of this stage is to link a given fragment to the docked chemical core at a specified position; the atoms provided in the previous step will be involved in the creation of a new single covalent bond. To achieve this, the coordinates of the hydrogens associated with these heavy atoms will be aligned, and the hydrogens will be subsequently deleted to produce the new bond (Figure 2A-2B). If any of the chosen heavy atoms have more than one hydrogen, the user can specify which one to use, controlling the resulting stereocenter, otherwise, the algorithm automatically selects the one that produces fewer clashes with the protein. Terminal heavy atoms other than hydrogen atoms are automatically replaced for a hydrogen.

In some cases, the addition of the fragment would lead to intramolecular clashes with the core of the molecule. In this scenario, rotations of 10 degrees along the axis of the new bond are successively performed until a more favorable position is located. If no position is found, the resolution of the rotamer is increased to 1 degree; the program stops if no position is found at this

resolution. At the end of this stage, the terminal atom of the initial core is replaced with the fragment to grow, obtaining the complete ligand. Additionally, the force field parameters (2005 OPLSAA)[36] of the ligands (initial and grown) are generated.

3 - Fragment Reduction. This stage reduces the parameters of the atoms in the fragment to be later grown dynamically within the binding site. The specific non-bonding terms being reduced are the next: σ Lennard-Jones potential (Van Der Waals radii), q (charge) of the electrostatic potential, and the r (equilibrium distance) of the bonding parameters. The reduction is applied differently for each parameter. The initial σ_o and q_o are computed following equation 1 and 2, respectively where σ is the Van Der Waals radii of the grown ligand, qH are the charges of the hydrogen that has been replaced from the protein-ligand complex, N is the number of atoms of the fragment, and L is the number of growing steps (GS) to be performed.

$$\sigma_o = \frac{\sigma}{L+1} \tag{1}$$

$$q_o = \frac{qH}{(L+1) \times N} \tag{2}$$

The computation of the initial bond distances (r_o) follows equation 3, where r is the grown ligand equilibrium distance, except for the initial distance involving the linking atoms, that is calculated with the equation 4, being rH the equilibrium distance of the bond between the atoms of the core and the hydrogen atom that has been replaced. The distance between the fragment and the core is increased in comparison with the bonds of the fragment to avoid intramolecular clashes at the beginning of the simulation. At the same time, the (resulting) charge of the hydrogen atom is spread and reduced proportionally to the number of GS into the different atoms of the fragment to simulate the electrostatics of the initial hydrogen. Notice that this charge is initially reduced by GS since larger charge values introduced artifacts in the initial sampling.

Angles and dihedrals are initially kept the same. A default value of 10 for GS is chosen, providing a good balance between a smooth and fast growing, and avoiding large artifacts in the H-like volume of the fragment produced by too low or too high *L*.

$$r_o = \frac{r}{L+1} \tag{3}$$

$$r_{link} = rH + \frac{r - rH}{L+1} \tag{4}$$

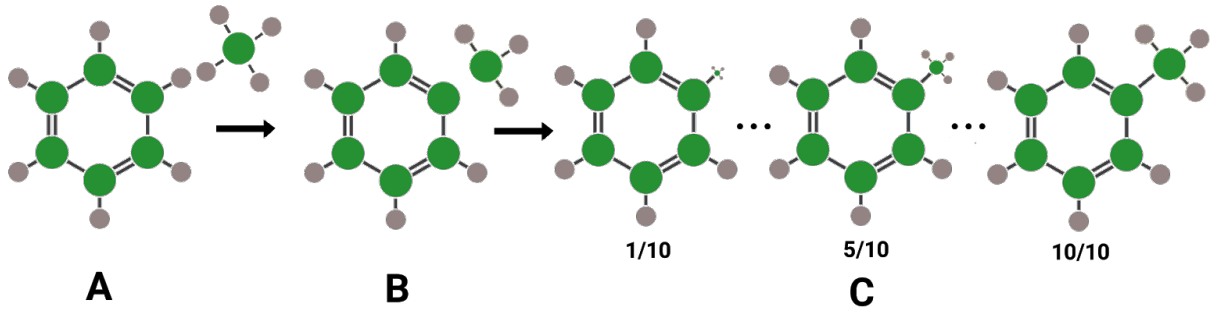

Figure 2. Growing of a methyl onto a benzene core. Carbon and hydrogen atoms are depicted in green and gray, respectively. Initially, (A) the fragment and the core are separated as two different entities; (B) the fragment is correctly placed by aligning and deleting the hydrogens that will be involved in the creation of the new covalent bond. Finally, (C) the fragment is miniaturized and ready to be grown along the simulation in 10 growing steps.

4 - Fragment growing. Starting from the concept introduced by AdaptivePELE, the new fragment is grown in a series of epochs, called growing steps (GS). At each GS, parallel MC simulations (47 simulations of 6 MC steps by default) are run after linearly increasing the parameters of the fragment following equation 5 (where *X* is *σ*, *q* and *r*, and *S* is the current GS). To increase the sampling, after each epoch all resulting structures are clustered using protein-ligand contact maps and the k-means algorithm. For each cluster, the structure with the best interaction energy is used as input for the next growing step. The interaction energy is computed

subtracting the energy of the protein and the ligand (both isolated) to the total energy of the system [E(AB) -E(A) -E(B)]. The energy is calculated with equation 6, having into account bonding, non-bonding and solvation energy terms. Thus, the fragment growing is an iterative process where the output structures of the current GS are the input for the next iteration. Figure 3 illustrates the process for growing a chlorophenyl fragment onto amino-indazole.

$$X_{step} = \lambda X \,;\, \lambda = \frac{1}{L - S + 1} \tag{5}$$

$$E = E_{bond} + E_{angle} + E_{torsion} + E_{improper\,torsion} + E_{vdw} + E_{ele} + \Delta G_{solv,\,pol} + G_{solv,\,npol} + E_{constraints} \tag{6}$$

5 - Sampling simulation and analysis. Once the ligand is completely grown, a longer PELE simulation is performed to score the grown molecule, by default involving 20 MC steps. In this, side-chain sampling is emphasized to thoroughly map the protein-ligand conformational space, to obtain all possible rearrangements of the complex. The score is computed as the mean of the 25 percent lowest values of interaction energies along with the simulation, to minimize the effects of outliers and false positives.

As a summary, the standard protocol to grow one fragment into a core with FragPELE consists of 10 GS, containing 47 independent MC simulations of 6 PELE steps each, and a final sampling simulation of 20 PELE steps. Every GS results on 5 different clusters to initialize the next iteration. Moreover, to restrict the core's perturbation, all simulations are configured with low translations, 0.05-0.10 Å, and rotations, 0.02-0.05 radians, and allowing only displacements within a box of 4 Å radius. In terms of computation time, growing one fragment takes approximately 1 hour on two Intel Xeon Platinum 8160 processors (2x24 processor units).

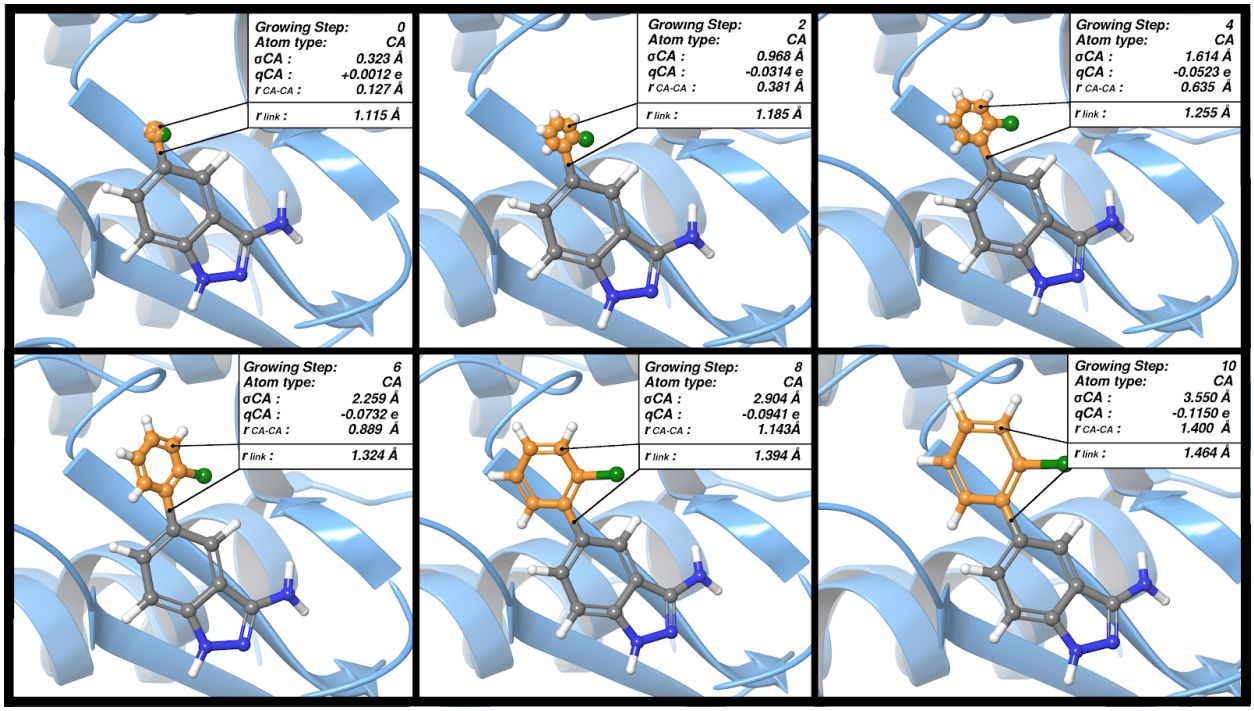

Figure 3. Example of a FragPELE simulation where a chlorophenyl fragment is grown onto an amino-indazole in 10 growing steps. Carbon and nitrogen atoms of the core are represented in gray and blue, respectively. Carbon and chlorine atoms of the fragment are represented in orange and green, respectively. For each panel, the structure and the values of the force field parameters are shown in the inset.

Benchmark preparation. Two main different tests have been carried out. First, a structural validation of the capacity of reproducing X-ray crystal structure poses was performed. Second, we evaluated the ability of FragPELE at successively growing fragments and scoring their affinity.

Structural benchmark. This benchmark aims to evaluate whether FragPELE can reproduce native poses of an X-ray crystal structure when growing in different scenarios. A total of 13 systems were analyzed, where four of them were focused on cryptic sub-pocket identification.

The first validation consisted of growing part of the ligand that has been previously removed from the original X-ray crystal structure. We have called this concept “self-growing”. Thus, simulations will start from the crystal with the remaining scaffold already placed in the binding site and the part of the ligand that has been previously removed will be grown to try to reproduce the native pose. For this, three different systems with available crystallographic data were chosen from the FEP+ Benchmark,[37] as all ligands were obtained from FBDD efforts. The chosen systems were:

1. Major Urinary Protein (MUP-I) in complex with sec-butyl-thiazoline (PDB code: 1I06).[38]
2. p38α kinase or mitogen-activated protein kinase 14 (MAPK14) co-crystallized with 3-(benzyloxy)pyridin-2-amine (PDB code: 1W7H).[39]
3. Bacterial DNA ligase in complex with an azaindazol (PDB code: 4CC6).[40]

All fragments were first manually removed from their respective X-rays by replacing them by a hydrogen atom (Table S1). Then, these fragments were grown again and the quality of the results was assessed by computing the heavy atom root mean square deviation (RMSD) of the core, fragment and the entire ligand against the initial crystal structure. p38 and DNA ligase were tested with and without structural water molecules; no structural water molecules were found in MUP-I due to the high hydrophobicity of the binding site.

The second structural test consisted of growing one or more fragments onto an X-ray scaffold to reproduce the interactions found in a second X-ray structure containing the entire ligand. We refer to this concept as “cross-growing”.

For this, four different systems were used, from which we had at least two crystal structures, one co-crystallized with a core ligand and a second one co-crystallized with a larger ligand, which could be generated by growing a fragment onto the core ligand. The heavy atom RMSD of

the core, fragment and fully grown ligand against the second X-ray structure were used to evaluate the quality of the results. Key ligand-protein interactions of the native structure were compared with the ones found in the structures retrieved by FragPELE, to evaluate the ability of the software at reproducing native interactions while growing.

The four systems used are, T4 Lysozyme, p38α kinase, the tyrosine-protein kinase JAK2, and beta-secretase 1 (BACE). Scaffold and fragments for all systems are shown in Table S2. Water molecules were deleted from all systems, except for the growing of BACE (4DJU to 4DJW), in which the waters were kept to reproduce an interaction between the fragment and the water molecules.

Finally, we tested the ability of FragPELE in sub-pocket identification. Cryptic sub-pockets are hidden cavities within a well-known binding site, which only open up when induced by the presence of particular R-groups. This benchmark evaluated whether FragPELE could reproduce the cryptic sub-pocket opened when growing the second-generation inhibitor lapatinib (PDB code: 1XKK)[41] from the first-generation inhibitor gefitinib (PDB code: 4WKQ) on the epidermal growth factor receptor (EGFR). As gefitinib had too many rotatable bonds to sample all possible ligand conformations in an acceptable amount of time, we decided to delete its solvent-exposed R-group, generally used to modify the ADME properties of the drug (Table S3), as this part is irrelevant for the present study. To evaluate results, we focused the analysis on two side chains lining the sub-pocket, M766 and F856. As it is seen in the crystal structure, the former must move aside to generate enough space to place the bulky fragment, and the latter is engaging the drug in a crucial pi interaction with the fluoro-phenyl ring.

In addition, we tested the reliability when growing a fragment that decreases the binding affinity of its precursor. Thus, FragPELE was tested at growing a small series of MAPK p38

inhibitors, starting from the inhibitor found in crystal structure 1A9U.[42] IC50s are available for all analogs (Table S4) but not their binding modes. The series was obtained from CHEMBL[43] (CHEMBL71403,[44] CHEMBL69929),[45] where we made sure the members have a wide range of experimental values. The system deals with a prototypical ATP-competitive kinase inhibitor that anchors a heterocycle to the central hinge residue (in this case M109) via an H-bond.[40] We hypothesize that the less potent ligands within the series have too bulky R-groups which cannot be accommodated. To verify this, structures retrieved from FragPELE simulations were compared to the initial pose of the only analog whose X-ray structure is available.

Growing and scoring benchmark. We evaluated FragPELE's performance at growing and scoring fragments in 5 systems on the FEP+ benchmark of Steinbrecher et al.:[37] T4 Lysozyme, DNA ligase, MUP-I, JAK-II, and p38. These systems were carefully chosen in order to have variability on the fragment size, binding site characteristics, and MW correlation. We removed those cases where the molecule was not amenable to R-group growing methodologies, such as molecules formed by single rings or alchemical transformations of heavy atoms of the core (Table S5 to S9). Standard Induced-Fit Glide calculations (with OPLS3 force field[46]) were run and our results were compared to FEP+, Glide SP docking and MM-GBSA, as provided in the benchmark paper.[37]

Set up. Structures with missing atoms in side-chains were corrected using the 3D builder of Maestro,[47] followed by energy minimization of these residues with the OPLS-2005 force field[36] and the implicit SGB solvent.[48] The same force field and solvent was used in FragPELE simulations. As stated, all systems were prepared with Schrödinger's Protein Preparation Wizard[34] including H-bond optimization using PROPKA[35] at pH 7.

RESULTS AND DISCUSSION

The method was validated in terms of generating good binding geometries (structural validation) and scoring congeneric series of ligands.

Structural Validation. RMSD of the self-growing test are shown in Table 1 (upper panel). All systems show a heavy atom RMSD values under 2 Å, retrieving native-like conformations for fragment and core. However, close attention must be paid to water molecules in hydrated sites, as the RMSD slightly increase when removing them. A clear example is shown in Figure S1, where the core and fragment RMSDs increase when water molecules are not present; the main reason being that the ligand tends to occupy the steric space left by the removed waters despite no further interactions are gained. Finally, results from the DNA ligase (4CC6) system, shown in Figure 4, illustrate that FragPELE is prone to recover native interactions of the complex, as the trifluoride of the fragment is found to rapidly interact with the surrounding hydrophobic residues, stabilizing this conformation along the simulation.

Cross-growing results are seen in Table 1 (bottom panel). All RMSD values excepting the p38 system fall below 2 Å. A closer inspection at the best structures for p38 (Figure 5) revealed that the naphthyl had been rotated almost 180 degrees, increasing the RMSD of the fragment. As depicted in Figure 5, the addition of the fragment causes a displacement of K53 towards a crucial pi-cation interaction with the ligand, and yields a native-like pose with 2.69 Å RMSD with respect to the crystal. Despite the new conformation of the system, K53 still conserves the specific lysine-glutamine lock of kinases.

Table 1. Results of the self-growing and cross-growing in terms of core, fragment and total heavy atom RMSD.

	Protein	PDB code(s)	Waters in simulation	RMSD core (Å)	RMSD fragment (Å)	Total RMSD (Å)	Figure
Self-growing	MUP-I	1I06	No	1.58	1.37	1.49	-
	p38	1W7H	No	2.53	2.46	2.51	S1-upper
			Yes	1.96	1.54	1.80	S1-bottom
	DNA ligase	4CC6	No	2.28	2.04	2.15	-
			Yes	1.23	1.59	1.46	4
Cross-growing	Lysozyme	181Ll to 184L	No	0.34	1.1	0.75	-
	p38	1W7H to 1WBW	No	1.38	3.46	2.69	5
	JAK-II	3E62 to 3E63	No	0.47	1.67	1.08	S2
	BACE	4DJU to 4DJV	No	0.79	0.96	0.84	-
		4DJU to 4DJW	Yes	1.02	0.41	0.92	S3
		4DJX to 4DJY	No	1	1.34	1.05	-

For the JAK-II system, in the growing of a phenyl onto the core of the crystal 3E62 (Figure S2), FragPELE was able to relocate the aspartic acid side-chain to accommodate the six-membered ring of the fragment with a 1 Å RMSD. Lastly, for the BACE system, in 4DJW we reproduced a crucial interaction between a structural water present in the crystal and the ligand (Figure S3). This interaction locks the conformation of the ligand and enthalpically favors the position of the explicit water.

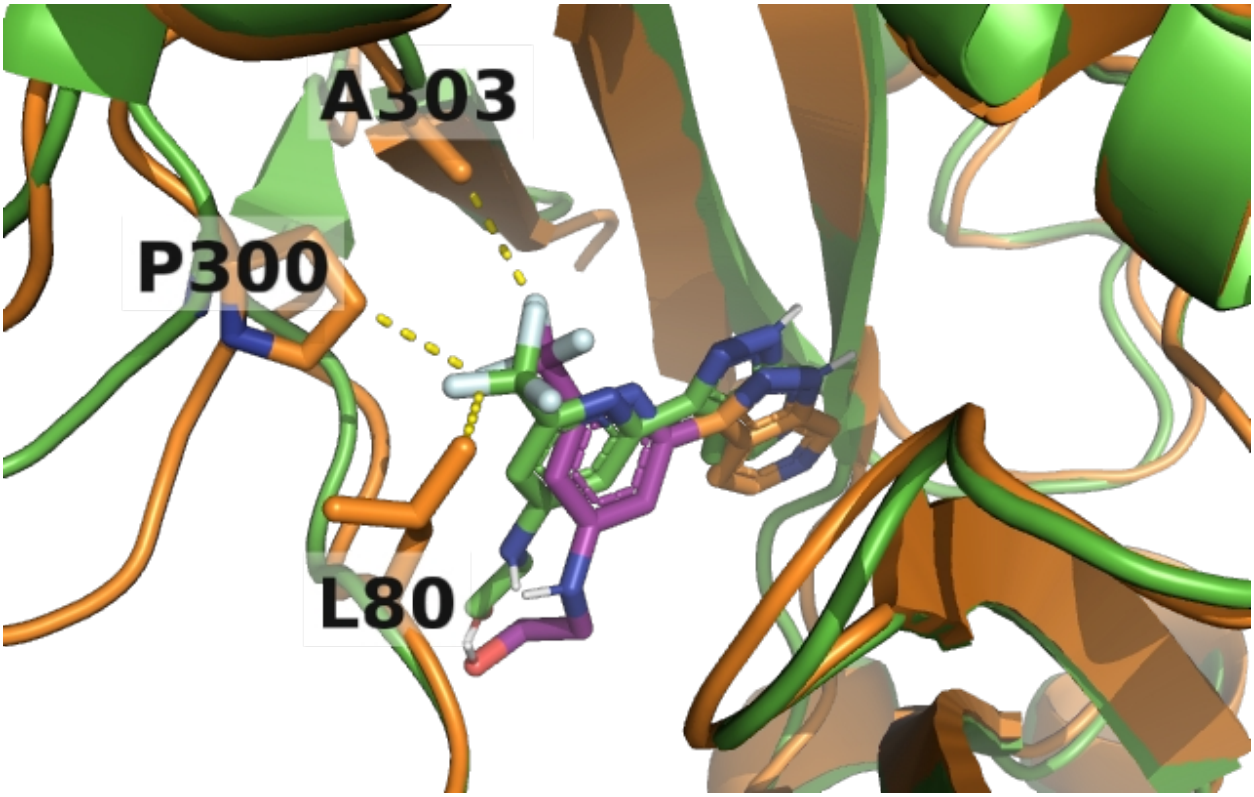

Figure 4. Growing of the fragment (purple) onto the core (orange) to reproduce PDB 4CC6 (green). Yellow dashed lines are highlighting the interactions between the surrounding residues of the receptor and the trifluoride of the fragment.

Regarding cryptic sub-pockets, we used the epidermal growth factor receptor (EGFR), a tyrosine kinase frequently overexpressed in many types of cancers.[49] Here, we focused on the development of lapatinib,[6] which was created by extending gefitinib[5] seeking to target a novel ligand-induced cavity next to the ATP binding site. FragPELE results are shown in Figure 6 where we observe how the ligand fragment pushes M766 away and concertedly moves F856 closer to the grown phenyl to produce a stabilizing pi-pi stack interaction. As it is seen in Figure S4, all amino acids that were originally lining close to the fragment in lapatinib X-ray are also present in the model. The pi-pi stack between F856 is not represented in the model because the

flip of 180º of the ring (Figure S5), which interposes the fluorine atom; however, centroid distances (5.7 Å) and angles (54º) are close to suitable conditions for this kind of interaction.

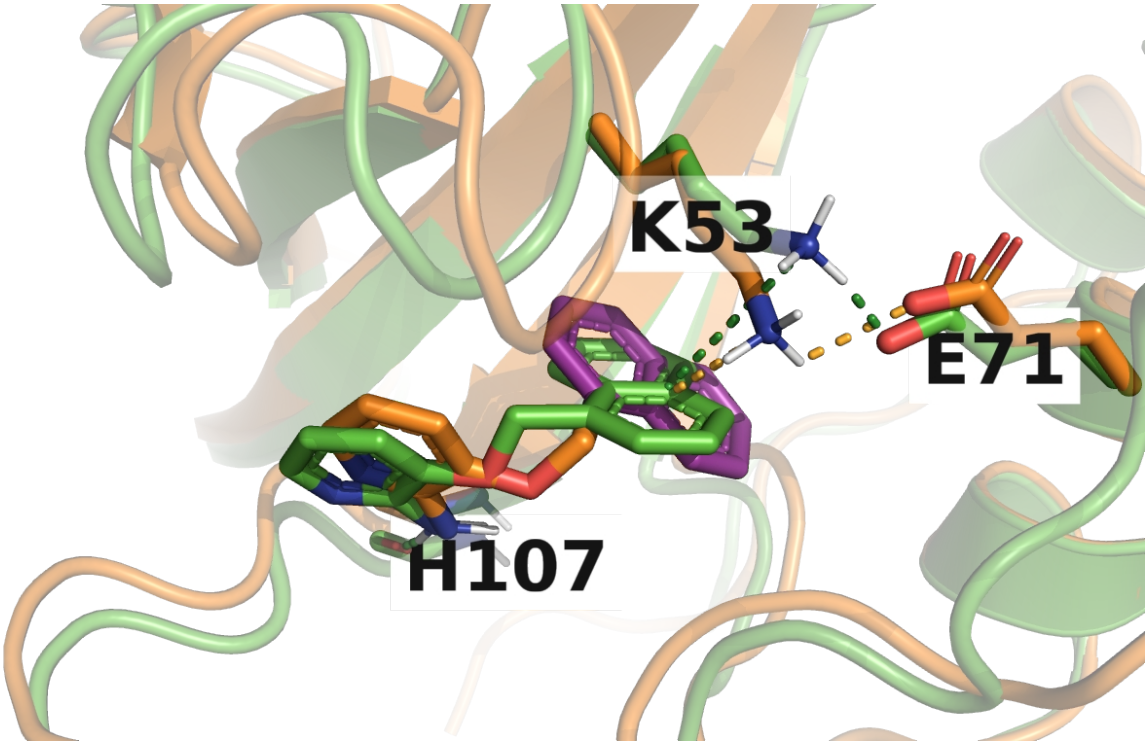

Figure 5. Growing of naphthyl (purple) and amino fragments from the core of 1W7H (orange) to reproduce the crystal structure 1WBW (in green). The addition of the naphthyl fragment moves K53 in order to allow the pi-cation, favoring the interaction between K53 and E71, which is present in the crystal. Subsequently, the addition of the amino fragment is creating a second interaction with H107. Thus, all native interactions present in 1WBW were recovered.

In order to verify FragPELE does not generate false positives when growing a bulky R-group, the binding mode of three p38 ATP-inhibitors with a common scaffold was predicted. Two of the fragments confer higher potency (low nM) whereas the other one presents lower affinity (low μM) than the reference compound due to a very bulky R-group (Table S4). Binding pose predictions are shown in Figure 7, where it is observed that the addition of cyclohexane promoted the displacement of the ligand outside the cavity (Figure 7, yellow model), losing the

canonical hinge interaction with M109 (a backbone interaction). This contact, still present in the other two ligands, is well known to account for a great part of the binding affinity, suggesting that the cyclohexane ring is too big to fit into the ATP binding site, probably causing a decrease of binding potency due to steric effects.

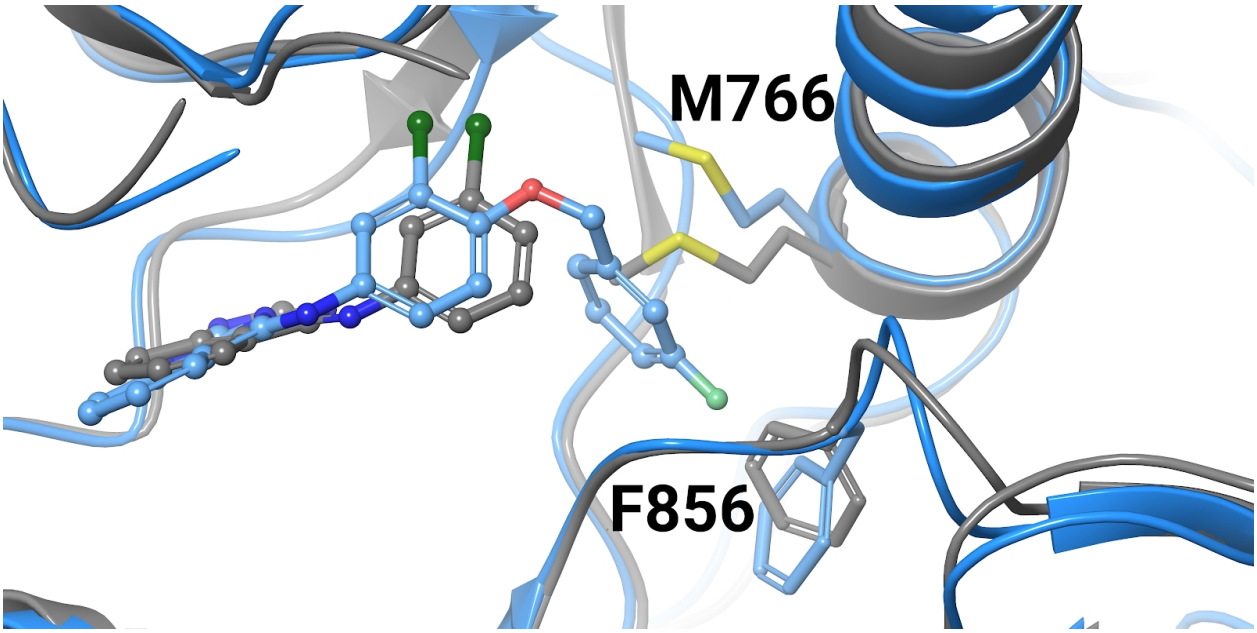

Figure 6. Cross-growing of lapatinib from gefitinib. Result (in blue) onto the initial protein-ligand complex (in gray). M766 moves aside allowing the fragment to grow and the F856 re-orients towards the fragment aromatic group.

Backbone RMSD was also analyzed for the different systems, please refer to the "Protein structure analysis" section of the Supplementary information. We have obtained low RMSD values for those systems with lower mobility (MUP-1, DNA ligase, BACE and lysozyme), and higher ones in kinases (p38 and JAK-II), which have more flexibility (Table S10). However, all protein RMSD were quite low (between 1.16 and 4.67Å) and we could consider that side chain rearrangements were fundamental to reproduce crystal structures.

Growing and scoring. The structural validation indicated that FragPELE is capable of correctly predicting the ligand-bound geometry within a binding site after R-group growth, even in cases where significant MW gains are involved. As a final step, we assessed whether the interaction energies generated for the grown R-groups could be used to rank the grown molecules in a H2L stage. PELE‘s interaction energy typically discriminates well ligands of similar size against the same target. Thus, we hypothesize that they may work when trying to relatively score ligands with a common structural core; where we expect differences in entropic terms to be small compared to the change in enthalpy. The chosen benchmark involves the FEP+ original study,[37] which allowed us to compare our technique with state of the art techniques such as Glide, MMGBSA, and FEP+. Hence, this benchmark only assessed the accuracy of FragPELE at growing and scoring fragments but also if the software falls midway between the accurate but expensive FEP and the cheap but sometimes inexact docking algorithms.

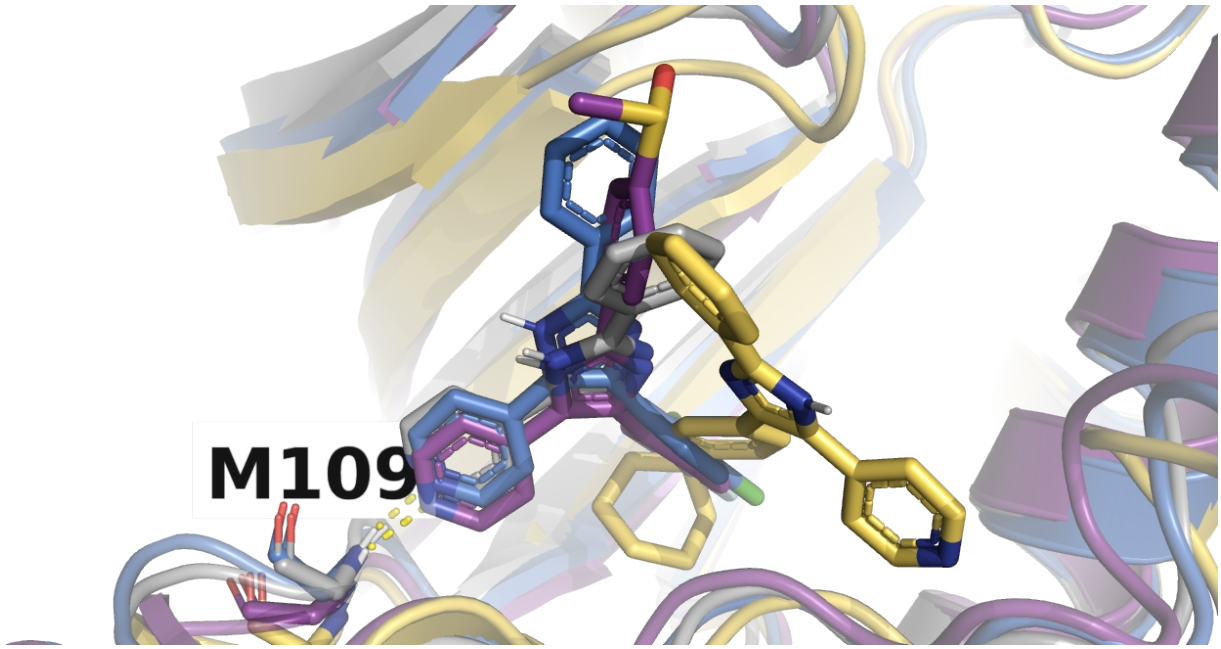

Figure 7. Growing of several fragments to create three different inhibitors of p38 type II. 1A9U (purple) and CHEMBL69929 (blue) correspond to binders with significantly higher affinity, and

CHEMBL71403 (yellow) as the lowest affinity. In gray it is represented the position of the core structure, derived from crystal 1A9U where we deleted the fluorine and methanesulfonyl groups. Notice that the models of 1A9U and CHEMBL69929 keep the canonical interaction with M109, contrary to CHEMBL71403.

A complete view of the results is shown in Table 2; individual correlation energy plots are shown in Figures S6-S10. The correlations obtained with FragPELE are only slightly worse than FEP+ for lysozyme, DNA ligase, and JAK-II systems. We obtain similar results as FEP+ in MUP-I and, surprisingly, outperform FEP+ in the p38 results. Moreover, our results are clearly better than Glide SP scores for almost all systems as the latter does not account for side-chain flexibility. However, for p38 and lysozyme, which have a low MW correlation, Glide (SP and Induced-Fit), and MMGBSA perform poorly, while FragPELE and FEP+ obtain good correlation values. Finally, for JAK-II (0.32 MW correlation) all methods seem to struggle, with only FEP+ achieving an acceptable correlation (0.64).

Table 2. FragPELE results overview and comparison to experimental data between different scoring approaches from FEP+ benchmark of Thomas Steinbrecher et.al.[37]

System	PDB code	R^2 FragPELE	R^2 FEP+	R^2 Glide SP*	R^2 Glide Induced-Fit	R^2 MMGBSA default*	R^2 MMGBSA flexible*	R^2 MW
Lysozyme	181L	0.64	0.79	0.32	0.28	0.40	0.3	0.32
DNA ligase	4CC5	0.88	0.98	0.36	0.75	0.01	0.36	0.92
MUP-I	1I06	0.96	0.94	0.92	0.84	0.86	0.75	0.93
JAK-II	3E62	0.48	0.64	0.50	0.19	0.50	0.21	0.32
p38	1W7H	0.87	0.69	0.09	0.50	0.01	0	0.63

* R^2 directly extracted from FEP+ benchmark.

Interestingly, FEP+ outperforms FragPELE in terms of accuracy when ranking systems where the MW correlation is lower than 0.5. However, both of them perform similarly when the gain of

further interactions (additional mass) between fragment and receptor accounts for the affinity change. This seems to indicate that enthalpic contributions might be described accurately enough with our simpler technique. We remind here that our score is simply the mean of the lowest 25% force field interaction energies (see Methods section); a significance source of error might come from the lack of explicit solvent and entropic effects (both accounted for in FEP techniques).

Regarding computing time, FragPELE takes an average of one hour per fragment on 48 Intel Xeon Platinum 8160 processor and can be run on any commodity CPU cluster. Therefore, its computational cost falls midway between FEP and docking, but still accounting for the dynamics of the ligand-protein system.

CONCLUSIONS

In this paper, we performed the first benchmarks of FragPELE as a method for dynamic fragment growing, where we tested its potential for predicting protein-bound geometries and for assessing the affinities of the grown ligands within congeneric series. Structural results show good correlations between crystallographic data and predicted structures, even in cross-growing tests where the grown derivatives imply significant geometric receptor accommodation. Predictions for some cases show a clear improvement when explicit water molecules are incorporated. For this reason, in prospective studies, the water position should be determined through the previous knowledge of the system (using, for example, specific software such as WaterMap).[50] Other limitations of this first version include the requirement of growing from a hydrogen atoms, cycle addition and core atoms transformations; we are working on solving all these issues in the next version. Regarding the size, there is no limit in the fragments being added, large fragments, however, might result in strong steric clashes and introduce unrealistic poses. Importantly, FragPELE seems promising at locating which areas of a binding site are

prone to opening up unexpected sub-pockets when probed with specific R-groups, which warrants further efforts to benchmark the technology. Interestingly, while FragPELE was not designed to provide free binding energies, its score was compared to FEP+, Glide SP, Glide Induced-Fit, and MMGBSA, clearly outperforming the three last techniques and providing a close answer to the significantly more computationally intensive FEP+ method.

In summary, the combination of our adaptive Monte Carlo sampling with a stepwise growing algorithm allows the ligand-receptor complex to rapidly adapt while exploring the most significant areas of the potential energy surface.

SUPPORTING INFORMATION

Tables showing the chemical compounds used in the benchmark, and extra figures of the models and diagrams to support the explanation of the results (PDF).

FragPELE code is available in GitHub: https://github.com/carlesperez94/frag_pele

CORRESPONDING AUTHORS

Victor Guallar: Tel.: +34 93 413 7727. E-mail: victor.guallar@bsc.es

ORCIDS

Carles P. Lopez: 0000-0002-2876-0947

Daniel Soler: 0000-0003-3274-2482

Robert Soliva: 0000-0003-0628-4342

Victor Guallar: 0000-0002-4580-1114

CONFLICT OF INTEREST

The authors declare the following competing financial interest(s): V.G. has a significant financial stake in, and is head of the Scientific Advisory Board of Nostrum Biodiscovery.

ACKNOWLEDGMENTS

This work was supported by CTQ2016-79138-R and the RTC-2017-6295-1 grants from the Spanish Government. Nostrum Biodiscovery is supported by Fundación Marcelino Botín (Mind the Gap) and CDTI (Neotec grant−EXP 00094141/SNEO-20161127) and wishes to thank support from BSC and IRB.

ABBREVIATIONS

BACE, beta-secretase 1; CPU, central processing unit; EGFR, epidermal growth factor receptor; FBDD, fragment-based drug design; FEP, free energy perturbation; GS, growing step; H2L, hit to lead; MAPK14, mitogen-activated protein kinase 14; MC, Monte Carlo; MD, molecular dynamics; MUP-I, major urinary protein 1; MW, molecular weight; PELE, protein energy landscape exploration; RMSD, root mean square deviation; RNN, recurrent neural networks; VAE, variational autoencoders.

REFERENCES

(1) Hughes, J. P.; Rees, S.; Kalindjian, S. B.; Philpott, K. L. Principles of Early Drug Discovery. *Br. J. Pharmacol.* **2011**, *162*, 1239–1249.

(2) Brown, D. G.; Boström, J. Where Do Recent Small Molecule Clinical Development Candidates Come From? *J. Med. Chem*. **2018**, 9442–9468.

(3) Murray, C. W.; Rees, D. C. The Rise of Fragment-Based Drug Discovery. *Nature Chemistry.* **2009**, 187–192.

(4) Erlanson, D. A.; Fesik, S. W.; Hubbard, R. E.; Jahnke, W.; Jhoti, H. Twenty Years on: The Impact of Fragments on Drug Discovery. *Nat. Rev. Drug Discov.* **2016**, *15*, 605–619.

(5) Rawluk, J.; Waller, C. F. Gefitinib. *Recent Results Cancer Res*. **2018**, *211*, 235–246.

(6) Voigtlaender, M.; Schneider-Merck, T.; Trepel, M. Lapatinib. *Recent Results Cancer Res.* **2018**, *211*, 19–44.

(7) Schrödinger. CombiGlide; **2019**.

(8) White, D.; Wilson, R. C. Generative Models for Chemical Structures. *J. Chem. Inf. Model.* **2010**, 1257–1274.

(9) Yuan, Y.; Pei, J.; Lai, L. LigBuilder 2: A Practical de Novo Drug Design Approach. *J. Chem. Inf. Model.* **2011**, *51*, 1083–1091.

(10) Durrant, J. D.; Lindert, S.; McCammon, J. A. AutoGrow 3.0: An Improved Algorithm for Chemically Tractable, Semi-Automated Protein Inhibitor Design. *J. Mol. Graph. Model.* **2013**, *44*, 104–112.

(11) Vinkers, H. M.; de Jonge, M. R.; Daeyaert, F. F. D.; Heeres, J.; Koymans, L. M. H.; van Lenthe, J. H.; Lewi, P. J.; Timmerman, H.; Van Aken, K.; Janssen, P. A. J. SYNOPSIS: SYNthesize and OPtimize System in Silico. *J. Med. Chem*. **2003**, *46*, 2765–2773.

(12) Sommer, K.; Flachsenberg, F.; Rarey, M. NAOMInext - Synthetically Feasible Fragment Growing in a Structure-Based Design Context. *Eur. J. Med. Chem.* **2019**, *163*, 747–762.

(13) Schneider, G.; Fechner, U. Computer-Based de Novo Design of Drug-like Molecules. *Nat. Rev. Drug Discov.* **2005**, *4*, 649–663.

(14) Douguet, D.; Munier-Lehmann, H.; Labesse, G.; Pochet, S. LEA3D: A Computer-Aided Ligand Design for Structure-Based Drug Design. *J. Med. Chem.* **2005**, *48*, 2457–2468.

(15) Dean, P. M.; Firth-Clark, S.; Harris, W.; Kirton, S. B.; Todorov, N. P. SkelGen: A General Tool for Structure-Based de Novo Ligand Design. *Expert Opin. Drug Discov*. **2006**, *1*, 179–189.

(16) Chéron, N.; Jasty, N.; Shakhnovich, E. I. OpenGrowth: An Automated and Rational Algorithm for Finding New Protein Ligands. *J. Med. Chem.* **2016**, *59*, 4171–4188.

(17) Merk, D.; Friedrich, L.; Grisoni, F.; Schneider, G. De Novo Design of Bioactive Small Molecules by Artificial Intelligence. *Mol. Inform*. **2018**, *37*

(18) Genheden, S.; Ryde, U. The MM/PBSA and MM/GBSA Methods to Estimate Ligand-Binding Affinities. *Expert Opin. Drug Discov.* **2015**, *10*, 449–461.

(19) Wang, L.; Wu, Y.; Deng, Y.; Kim, B.; Pierce, L.; Krilov, G.; Lupyan, D.; Robinson, S.; Dahlgren, M. K.; Greenwood, J. Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field. *J. Am. Chem. Soc.* **2015**, *137*, 2695–2703.

(20) Cournia, Z.; Allen, B.; Sherman, W. Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations. *J. Chem. Inf. Model*. **2017**, 2911–2937.

(21) Pérez-Benito, L.; Casajuana-Martin, N.; Jiménez-Rosés, M.; van Vlijmen, H.; Tresadern, G. Predicting Activity Cliffs with Free-Energy Perturbation. *J. Chem. Theory Comput.* **2019**, 1884–1895.

(22) Jorgensen, W. L. Efficient Drug Lead Discovery and Optimization. *Acc. Chem. Res.* **2009**, *42*, 724–733.

(23) Gill, S. C.; Lim, N. M.; Grinaway, P. B.; Rustenburg, A. S.; Fass, J.; Ross, G. A.; Chodera, J. D.; Mobley, D. L. Binding Modes of Ligands Using Enhanced Sampling (BLUES): Rapid Decorrelation of Ligand Binding Modes via Nonequilibrium Candidate Monte Carlo. *J. Phys. Chem. B* **2018**, *122*, 5579–5598.

(24) Borrelli, K. W.; Vitalis, A.; Alcantara, R.; Guallar, V. PELE: Protein Energy Landscape Exploration. A Novel Monte Carlo Based Technique. *J. Chem. Theory Comput.* **2005**, *1*, 1304–1311.

(25) Carlson, H. A.; Smith, R. D.; Damm-Ganamet, K. L.; Stuckey, J. A.; Ahmed, A.; Convery, M. A.; Somers, D. O.; Kranz, M.; Elkins, P. A.; Cui, G. CSAR 2014: A Benchmark Exercise Using Unpublished Data from Pharma. *J. Chem. Inf. Model*. **2016**, *56*, 1063–1077.

(26) Grebner, C.; Lecina, D.; Gil, V.; Ulander, J.; Hansson, P.; Dellsen, A.; Tyrchan, C.; Edman, K.; Hogner, A.; Guallar, V. Exploring Binding Mechanisms in Nuclear Hormone Receptors by Monte Carlo and X-Ray-Derived Motions. *Biophys. J.* **2017**, *112*, 1147–1156.

(27) Kotev, M.; Pascual, R.; Almansa, C.; Guallar, V.; Soliva, R. Pushing the Limits of Computational Structure-Based Drug Design with a Cryo-EM Structure: The Ca2+ Channel α2δ-1 Subunit as a Test Case. J. Chem. Inf. Model. **2018**, *58*, 1707–1715.

(28) Takahashi, R.; Gil, V. A.; Guallar, V. Monte Carlo Free Ligand Diffusion with Markov State Model Analysis and Absolute Binding Free Energy Calculations. *J. Chem. Theory Comput.* **2014**, *10*, 282–288.

(29) Gilabert, J. F.; Grebner, C.; Soler, D.; Lecina, D.; Municoy, M.; Gracia, O.; Soliva, R.; Packer, M.; Hughes, S.; Tyrchan, C. PELE-MSM: A Monte Carlo Based Protocol for the Estimation of Absolute Binding Free Energies. *J. Chem. Theory Comput.* **2019**.

(30) Lecina, D.; Gilabert, J. F.; Guallar, V. Adaptive Simulations, towards Interactive Protein-Ligand Modeling. *Sci. Rep*. **2017**, *7*, 8466.

(31) Gilabert, J. F.; Lecina, D.; Estrada, J.; Guallar, V. Monte Carlo Techniques for Drug Design: The Success Case of PELE. In *Biomolecular Simulations in Structure-Based Drug Discovery*; Gervasio, F. L., Spiwok, V., Eds.; Methods and Principles in Medicinal Chemistry; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, **2018**; pp 87–103.

(32) Cock, P. J. A.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B. Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics. *Bioinformatics* **2009**, *25*, 1422–1423.

(33) Bakan, A.; Meireles, L. M.; Bahar, I. ProDy: Protein Dynamics Inferred from Theory and Experiments. *Bioinformatics* **2011**, *27*, 1575–1577.

(34) Sastry, G. M.; Adzhigirey, M.; Day, T.; Annabhimoju, R.; Sherman, W. Protein and Ligand Preparation: Parameters, Protocols, and Influence on Virtual Screening Enrichments. J. Comput. Aided Mol. Des. **2013**, *27*, 221–234.

(35) Olsson, M. H. M.; Søndergaard, C. R.; Rostkowski, M.; Jensen, J. H. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. *J. Chem. Theory Comput*. **2011**, *7*, 525–537.

(36) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. J. *Am. Chem. Soc*. **1996**, *118*, 11225–11236.

(37) Steinbrecher, T. B.; Dahlgren, M.; Cappel, D.; Lin, T.; Wang, L.; Krilov, G.; Abel, R.; Friesner, R.; Sherman, W. Accurate Binding Free Energy Predictions in Fragment Optimization. *J. Chem. Inf. Model.* **2015**, *55*, 2411–2420.

(38) Timm, D. E.; Baker, L. J.; Mueller, H.; Zidek, L.; Novotny, M. V. Structural Basis of Pheromone Binding to Mouse Major Urinary Protein (MUP-I). *Protein Sci.* **2001**, *10*, 997–1004.

(39) Hartshorn, M. J.; Murray, C. W.; Cleasby, A.; Frederickson, M.; Tickle, I. J.; Jhoti, H. Fragment-Based Lead Discovery Using X-Ray Crystallography. *J. Med. Chem.* **2005**, *48*, 403–413.

(40) Howard, S.; Amin, N.; Benowitz, A. B.; Chiarparin, E.; Cui, H.; Deng, X.; Heightman, T. D.; Holmes, D. J.; Hopkins, A.; Huang, J. Fragment-Based Discovery of 6-Azaindazoles as Inhibitors of Bacterial DNA Ligase. ACS Med. *Chem. Lett.* **2013**, *4*, 1208–1212.

(41) Wood, E. R.; Truesdale, A. T.; McDonald, O. B.; Yuan, D.; Hassell, A.; Dickerson, S. H.; Ellis, B.; Pennisi, C.; Horne, E.; Lackey, K. A Unique Structure for Epidermal Growth Factor Receptor Bound to GW572016 (Lapatinib): Relationships among Protein Conformation, Inhibitor off-Rate, and Receptor Activity in Tumor Cells. *Cancer Res.* **2004**, *64*, 6652–6659.

(42) Wang, Z.; Canagarajah, B. J.; Boehm, J. C.; Kassisà, S.; Cobb, M. H.; Young, P. R.; Abdel-Meguid, S.; Adams, J. L.; Goldsmith, E. J. Structural Basis of Inhibitor Selectivity in MAP Kinases. *Structure* **1998**, *6*, 1117–1128.

(43) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40*, D1100–D1107.

(44) Chang, L. L.; Sidler, K. L.; Cascieri, M. A.; de Laszlo, S.; Koch, G.; Li, B.; MacCoss, M.; Mantlo, N.; O'Keefe, S.; Pang, M. Substituted Imidazoles as Glucagon Receptor Antagonists. *Bioorg. Med. Chem. Lett.* **2001**, *11*, 2549–2553.

(45) Liverton, N. J.; Butcher, J. W.; Claiborne, C. F.; Claremon, D. A.; Libby, B. E.; Nguyen, K. T.; Pitzenberger, S. M.; Selnick, H. G.; Smith, G. R.; Tebben, A. Design and Synthesis of Potent, Selective, and Orally Bioavailable Tetrasubstituted Imidazole Inhibitors of p38 Mitogen-Activated Protein Kinase. *J. Med. Chem.* **1999**, *42* , 2180–2190.

(46) Harder, E.; Damm, W.; Maple, J.; Wu, C.; Reboul, M.; Xiang, J. Y.; Wang, L.; Lupyan, D.; Dahlgren, M. K.; Knight, J. L. OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. *J. Chem. Theory Comput.* **2016**, *12*, 281–296.

(47) Schrödinger. Maestro; **2018**.

(48) Gallicchio, E.; Zhang, L. Y.; Levy, R. M. The SGB/NP Hydration Free Energy Model Based on the Surface Generalized Born Solvent Reaction Field and Novel Nonpolar Hydration Free Energy Estimators. *J. Comput. Chem*. **2002**, *23*, 517–529.

(49) Expression of EGFR in cancer - Summary - The Human Protein Atlas https://www.proteinatlas.org/ENSG00000146648-EGFR/pathology (accessed Jul 22, 2019).

(50) Abel, R.; Young, T.; Farid, R.; Berne, B. J.; Friesner, R. A. Role of the Active-Site Solvent in the Thermodynamics of Factor Xa Ligand Binding. *J. Am. Chem. Soc.* **2008**, *130*, 2817–2831.

For Table of Contents Use Only